codex alimentarius commission

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS

WORLD HEALTH ORGANIZATION

JOINT OFFICE: Viale delle Terme di Caracalla 00100 ROME Tel: 39 06 57051 www.codexalimentarius.net Email: codex@fao.org Facsimile: 39 06 5705 4593

CX 4/50 **CL 2005/44-MAS**
August 2005

TO: Codex Contact Points
Interested International Organizations

FROM: Secretary, Codex Alimentarius Commission
Joint FAO/WHO Food Standards Programme
FAO, 00100 Rome, Italy

SUBJECT: **Draft Guidelines for Evaluating Acceptable Methods of Analysis**

DEADLINE: **10 January 2006**

COMMENTS: To:

Secretary
Codex Alimentarius Commission
Joint FAO/WHO Food Standards Programme – FAO
Viale delle Terme di Caracalla
00100 Rome, Italy
Fax: +39 (06) 5705 4593
E-mail: codex@fao.org

Copy to:

Dr. Mária Váradi,
Central Food Research Institute (KÉKI),
H-1022 Budapest, Herman Ottó út 15
Fax No., +361.212.9853 & 361.355.8928
E-mail: m.varadi@cfri.hu

Background

1) The 27th Session of the Commission adopted the proposed draft Guidelines for Evaluating Acceptable Methods of Analysis[1] at Step 5 and comments were asked by the CL 2004/36-GEN at Step 6.

2) The 28th Session of CCMAS discussed the text of the draft Guidelines Section by Section[2]. Delegations made a number of general comments, and also commented on Scope and Requirements.

3) The Committee noted that additional comments provided by Member Governments required careful consideration and agreed to establish a Working Group led by New Zealand[3] which would work electronically in order to revise the document, preferably short, taking into account the discussion and written comments submitted at the current session.

4) The Working Group has reviewed all the comments, and has revised the Guidelines (see Appendix).

5) A revised format is used that separates out the various aspects of evaluation, and allows for the necessary linkages to related documents. Some additional points are included in the various sections in order to clarify the intention of the guidelines. The revised outline is:

Scope

Objectives

Requirements

---

1 ALINORM 04/27/23, Appendix V.

2 ALINORM 05/28/23, paras 8-20.

3 Argentina, Australia, Austria, Brazil, Dominica, European Community, Honduras, Japan, Republic of Korea, Netherlands, South Africa, Spain, United Kingdom, United States of America

[Definitions]

Annex A: Estimation of characteristics

Annex B: Conditions for Acceptance of Methods

Annex C: Examples of use of methods described in Annex A

6) A variety of changes have been made to Scope and Requirements, and a new section, Objectives, has been added. These changes have been made in response to comments, and they are explained in a column alongside the revised guidelines. The explanation is not intended to be part of the guidelines.

7) CCMAS discussed whether definitions should be included in the Guidelines. The Committee noted that the deletion of definitions by leaving a reference to Procedural Manual could ensure consistency in their use, however, in case of their amendments, it would be necessary to revise the document to make sure that amended definitions were appropriate for the purpose of the Guidelines. The definitions and related text have therefore been placed in square brackets for further consideration of whether they should be included in this document or elsewhere. During the development of the guidelines it is useful to have the definition alongside the section on estimation of the particular characteristic in Annex A. To avoid duplication, in the meantime these are not included in the Definitions section, which therefore only contains related definitions. The Working Group notes that analytical terminology is currently being revised by an electronic Working Group[4]. If definitions are retained in the Guidelines, it is recommended that they should be updated from this work.

8) Annex A has been substantially revised, and explanatory notes are included.

9) Annex B is new material that describes conditions under which a candidate method may be accepted as a general replacement for a standard method in judging product compliance after a method validation exercise. Explanatory notes for Annex B are included.

10) At CCMAS 28, the Delegation of the United States proposed to add a few examples on how to apply the Guidelines in a step by step manner for evaluating the acceptability of a specific analytical method. Annex C includes examples to illustrate the methods of calculation involved in the analysis of a trial according to the methods suggested in Annex A.

11) The Working Group also noted that another CCMAS Working Group is drafting a descriptive version of a document on conversion of the validation data of the methods for trace elements into criteria[5]. It is recommended that CCMAS should consider including the principles and procedures from this work in Annex B, and the examples in Annex C

The Proposed Draft Guidelines are hereby circulated for comments at Step 6 and will be considered by the 27th Session of the Committee on Methods of Analysis and Sampling, Budapest, Hungary, 15-19 May 2006.

Governments and international organizations wishing to provide comments should do so in writing, preferably by email, to the above addresses **before 10 January 2006**.

---

4 ALINORM 05/28/23, paras 43-51.

5 ALINORM 05/28/23, paras 93-99.

Appendix

DRAFT GUIDELINES FOR EVALUATING ACCEPTABLE METHODS OF ANALYSIS

Contents

Draft Guidelines	Explanation of changes
Black text = Wording from ALINORM 04/27/3, App. 3 Shaded text = Wording proposed by the Working Group	
SCOPE	
1. These guidelines provide a framework for evaluating acceptable methods of analysis.	The criteria approach is mentioned in the Requirements section.
2. The guidelines apply to methods that may be used for control, inspection or regulatory purposes in relation to import and export of foods.	This paragraph is added to clarify the types of laboratories to which the guidelines apply, using terminology from the Procedural Manual. A phrase is added to clarify the context in which the methods are used (Chile).
3. The guidelines specify criteria which methods must satisfy to be used as Type III methods. Some of the considerations may also apply to defining methods (Type I).	This paragraph is added to clarify the types of methods to which the guidelines apply.
4. The guidelines will not be applicable in some cases, for example where methods are not available in the public domain or where a method is being developed for a new analyte.	This paragraph is added to indicate that the guidelines will not apply in certain cases.
OBJECTIVES	
5. These guidelines ~~are intended to assist countries in the application of requirements for trade in foodstuffs~~ provide a scientific basis for the selection and acceptance of analytical methods to be used in assessments of product in order to protect the consumer and to facilitate fair trade.	This paragraph has been moved from the Scope (Chile). Phrasing is added to clarify the objective of the guidelines (New Zealand).
6. The guidelines are intended to allow more flexibility, through the development of appropriate criteria for methods as the basis for their acceptance.	This paragraph is added to clarify the objective of the guidelines.
REQUIREMENTS	
7. Acceptance of a method consists of the steps: (a) estimation of the performance characteristics of the method, (b) judgement of the method based on its performance characteristics and its fitness for purpose, and (c) formalizing acceptance of the method.	This paragraph is added to clarify the steps that are involved in acceptance of methods.
Estimation of the performance characteristics of a method	
8. Laboratories involved in the evaluation should comply with the Codex Guidelines for the Assessment of the Competence of Testing Laboratories Involved in the Import and Export Control of Foods (CAC/GL 27-1997).	This paragraph has been moved from the Scope. The title of the document has been corrected.
9. The following performance characteristics of the candidate method should be estimated ~~assessed~~ as appropriate ~~against the following criteria by laboratories involved in the import and export~~	This paragraph is reworded to clarify the intention of this step.

~~control of foods~~:

- ~~Accuracy~~

Accuracy is covered by the more fundamental characteristics of bias and precision (New Zealand)

- bias

This characteristic should be included (New Zealand).

- sensitivity.
- linearity
- precision (repeatability, reproducibility, and reproducibility net of repeatability)

To maintain consistency with Annex A (Brazil and Japan)

- limit[s] of detection[and quantification]

Some members of the WG consider that LoQ should not be included as it is not well defined and would depend upon the application. Others consider it should be defined and retained as a method performance characteristic.

- applicability (analytes, matrix, concentration range and preference given to 'general' methods)

"Analytes" added for agreement with Annex A (Brazil)

- recovery
- ruggedness (robustness)

This characteristic should be included (New Zealand).

- selectivity (interference effects etc.)

10. [The definitions of these characteristics ~~are given below, and approaches to their estimation~~ are in ….]

This paragraph is placed in square brackets for further discussion on whether definitions should be included in this document.

11. To the extent possible, the method's characteristics and the error associated with them should be estimated in a method performance study, conducted as recommended in Codex Food Control Laboratory Management: Recommendations (CAC/GL 28-1995, Rev.1-1997) The data from the method performance study should be analysed as described in Annex A. An adequate method description and verifiable performance data should be available for peer review.

This paragraph is added to reference the requirements for method performance studies.

12. In the case of single laboratory validation, there is lower confidence in the general applicability of the resulting estimates of statistical parameters than is available from interlaboratory studies.

13. Some characteristics such as precision and limit of detection can be also be applied in the case of defining methods (Type I).

This paragraph is added to clarify the considerations that can apply to Type I methods.

Judgement of the method based its performance characteristics and its fitness for purpose

14. Criteria should be established, involving relevant performance characteristics, for judging the acceptability of methods. The criteria may be specified as a requirement.

These paragraphs are added to describe the steps of judging acceptance.

15. Criteria should take account of:

- The performance characteristics of existing accepted methods. An example of this approach is given in Annex C;
- Importing country's requirements or requirements by the consumer;
- Specifications in food standards; or
- Fitness for purpose considerations in relation to the intended use of the results to judge conformity.

16. Assessment of fitness-for-purpose criteria may suggest a need to alter the method of judging conformity.

17. A candidate method may be accepted for general use if it satisfies the conditions outlined in Annex B.

Formalising acceptance of the method

18. Methods may be formally accepted in the country by:

- Approving specific methods that satisfy criteria, or
- Identifying or developing methods that satisfy specified criteria (see 14).

This paragraph is added to highlight that flexibility may be allowed by specifying criteria for methods.

19. When a method is accepted, the description of the method, estimates of the relevant characteristics and the demonstration that the criteria have been met should be formally documented.

This paragraph is added to describe the step of acceptance.

20. The scope of the acceptance should be determined by the range of experimental conditions under which the method has been tested. For example, in cases where performance characteristics have been estimated in fewer laboratories than specified in Annex A, the acceptance may be restricted.

21. If a method of analysis has been endorsed by Codex, then preference should be given to using that procedure.

This paragraph has been moved from the Scope.

[DEFINITIONS

This section is placed in square brackets for further discussion on whether definitions should be included in this document. CCMAS agreed that definitions would be reviewed by a working group.

Applicability

See Annex A

Bias

See Annex A

Fitness for Purpose

{IUPAC Harmonised Guidelines for Internal Quality Control in Analytical Laboratories}

Degree to which data produced by a measurement process enables a user to make technically and administratively correct decisions for a stated purpose.

Limit of Detection

See Annex A

Limit of Quantification

See Annex A

Linearity

See Annex A

Method validation

(to be developed)

Precision

See Annex A

Recovery

See Annex A

Repeatability [Reproducibility]:

This definition is taken from CX/MAS 05/26/6-Add. 2.

{ISO 3534-2}

Precision under repeatability [reproducibility] conditions.

Repeatability conditions

{ISO 3534-2}

Observation conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in the same test or measuring facility by the same operator using the same equipment within short intervals of time.

This definition is taken from CX/MAS 05/26/6-Add. 2.

Note:

Repeatability conditions include

the same measurement procedure or test procedure

the same observer

the measuring or test equipment used under the same conditions

the same location

repetition over a short period of time.

Repeatability [Reproducibility] Limit

{ISO 3534-2}

The value less than or equal to which the absolute difference between two final values each of them representing a series of test results or measurement results obtained under repeatability [reproducibility] conditions is expected to be with a specified probability of 95%.

This definition is taken from CX/MAS 05/26/6-Add. 2.

Notes:

1. The symbol used is r [R]. {ISO 3534-2}
2. When examining two single test results obtained under repeatability [reproducibility] conditions, the comparison should be made with the repeatability [reproducibility] limit, r [R] = 2.8sr [R]. {ISO 5725-6, 4.1.4}

Repeatability [Reproducibility] Standard Deviation

{ISO 3534-2}

The standard deviation of test results obtained under repeatability [reproducibility] conditions.

This definition is taken from CX/MAS 05/26/6-Add. 2. New definition of within-laboratory reproducibility standard deviation added (European Community).

Within-laboratory reproducibility standard deviation:

The standard deviation of test results obtained under within-laboratory reproducibility conditions.

Notes:

1. It is a measure of the dispersion of the distribution of test results under repeatability / reproducibility / within-laboratory reproducibility conditions.
2. Similarly “repeatability / reproducibility / within-laboratory reproducibility variance” and “repeatability / reproducibility / within-laboratory reproducibility coefficient of variation” could be defined and used as measures of the dispersion of test results under repeatability / reproducibility / within-

laboratory reproducibility conditions.

Standard deviation for reproducibility net of repeatability

The standard deviation $\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$ where σ_R is the reproducibility standard deviation and σ_r is the repeatability standard deviation.

New definition.

Note:
Knowledge of this standard deviation is necessary to assess the measurement error to which a sample mean is subject.

Reproducibility conditions

{ISO 3534-2}

Observation conditions where independent test/measurement results are obtained with the same method on identical test/measurement items in different test or measurement facilities with different operators using different equipment.

This definition is taken from CX/MAS 05/26/6-Add. 2. The note is from ALINORM 04/27/23, Appendix V.

Note:

When different methods give test results that do not differ significantly, or when different methods are permitted by the design of the experiment, as in a proficiency study or a material-certification study for the establishment of a consensus value of a reference material, the term “reproducibility” may be applied to the resulting parameters. The conditions must be explicitly stated.

Ruggedness (Robustness)

See Annex A

Selectivity

See Annex A

Sensitivity

See Annex A

Single laboratory validation

(to be developed)

Within-laboratory reproducibility (wR)

Precision under within-laboratory reproducibility conditions.

New definition (European Community).

Within-laboratory reproducibility conditions

Conditions where test results are obtained with the same method, on different test items in the same laboratory by different operators using different or the same equipment.]

New definition (European Community).

ANNEX A: ESTIMATION OF CHARACTERISTICS

THE "BLACK BOX" APPROACH

Because these guidelines are intended to cover a wide range of methods, a "black box" approach is used. A method is considered as a black box. A sample of material is put into the black box, and an estimate of concentration emerges. The inner workings of the method are not considered: only the relationship between what goes in and what comes out is relevant.

The output of a method may be essentially qualitative, for example, a yes/no response to the presence or absence of an organism. The considerations involved in the validation of such a method are substantially different from those when the response is quantitative, and such methods are not covered in these guidelines.

In this Annex, the "response" is taken to be the estimate of concentration emerging from the black box, rather than some intermediate quantity that is subsequently converted to such an estimate, for example by means of a calibration curve. In particular, sensitivity and linearity is considered in relation to the change in estimated concentration (output) for a given change in true concentration (input.)

Note

In some cases, the response would more appropriately be considered to be some function of concentration, such as the logarithm, that has behaviour that is more easily described.

TRIAL DESIGN

A method validation trial usually involves the presentation of samples of a number of homogeneous materials to each of a set of participating laboratories. Each material is presented to each laboratory.

The different parameters to be estimated during a validation trial fall into two groups. For estimating overall bias, reproducibility and limits of detection it is desirable that the samples should be analysed, within each laboratory, under conditions that are as varied as possible, whereas for the estimation of sensitivity and non-linearity the analysis of groups of samples under repeatability conditions is desirable.

For bias and reproducibility, the most efficient design, from a statistical point of view, is one where each sample is analysed in a different run. This is the least efficient design for estimating sensitivity. Consequently, if resources are limited, it may be desirable to split the validation into two parts. In one part sensitivity could be estimated and non-linearity investigated using a relatively small number of laboratories and runs, but with several samples per run. In another a large number of laboratories would be given a number of samples, but with the requirement to analyse the samples over a number of different runs under within-laboratory reproducibility conditions. For this second part, a minimum of eight laboratories is normally required (e.g. IUPAC Protocol) and if fewer laboratories are used the scope of the resulting method performance parameters may be restricted, for example to a particular laboratory or group of laboratories.

For the first part, this minimum could be reduced, mainly as a concession to practicality, but also on the grounds that sensitivity and non-linearity seem less likely to vary significantly between laboratories than overall bias does.

The range of concentrations used should extend over the range of concentrations for which the performance parameters are required and include at least five values (excluding duplicates) reasonably spread throughout the range. For the use of the test for non-linearity suggested below, at least ten values would be desirable.

REFERENCE VALUES

For the estimation of some performance parameters, particularly those involving bias, it is necessary that these samples should be of known concentration. Depending on the type of material being

studied, this may be achieved in various ways:

1. Artificial samples could be constructed with accurately known concentrations.

2. Reference samples could be used. In this and the following case the uncertainty attached to the reference values could create problems in estimating some performance parameters with sufficient accuracy.

3. Samples of the materials could be analysed by a reference method as well as by the candidate method at participating laboratories.

The considerations involved in estimating performance parameters may vary somewhat according to which of these situations applies.

BIAS

[Definition (ISO 3534-2)

The difference between the expectation of the test result or measurement result and the true value.

Notes:

1. Bias is the total systematic error as contrasted to random error. There may be one or more systematic error components contributing to bias. A larger systematic difference from the accepted reference value is reflected by a larger bias value. {ISO 3534-1}

2. The bias of a measuring instrument is normally estimated by averaging the error of indication over the appropriate number of repeated measurements. The error indication is the: "indication of a measuring instrument minus a true value of the corresponding input quantity"

3. In practice the accepted reference value is substituted for the true value

4. Expectation is the expected value of a random variable, e.g. assigned value or long-term average {ISO 5725-1}

REFERENCE:

ISO Draft Standard 3534-2: Vocabulary and Symbols Part 2: Applied Statistics, ISO, Geneva, 2004]

General comments on bias

Although individual estimates of bias may be made for each material presented, it is often more useful to summarize this information according to the following scheme:

Overall or mid-range bias

Uniformity of bias (method sensitivity)

Non-linearity.

These are discussed in separate subsections below

Estimates of the overall bias and of the method sensitivity are normally approximately statistically independent. Assuming this to be the case, and that no evidence of non-linearity has been found, the bias at any concentration within the range, together with a standard deviation expressing its uncertainty of estimation can then be estimated by

$$b(x) = b(x_0) + (s-1)(x - x_0)\,.$$

Here $b(x_0)$ is the estimated overall method bias, calculated as described above, and x_0 is the concentration at which it is assumed to apply. $b(x)$ is the bias at concentration *x* and *s* the estimated sensitivity.

This estimate of bias has the estimated standard error

$$s_{b(x)} = \sqrt{s_{b(x_0)}^2 + (x - x_0)^2 s_s^2},$$

where s_s is the standard error of the estimate of sensitivity and $s_{b(x_0)}$ is the standard error of the estimated bias.

If statistically significant evidence of non-linearity is found, this procedure will not be appropriate. Independent estimates of bias within various concentration ranges, using the only the samples whose values fall within those ranges, will be necessary. These estimates, with their standard errors, may be calculated using the method described in the section on overall bias below. This will also be necessary if estimates of sensitivity are not available, for example in cases where sufficiently precise reference values are not available, as discussed in the section on sensitivity.

Estimation of Overall Bias

It is desirable that estimates of overall bias, like estimates of reproducibility, should be based on a trial involving as many laboratories as possible.

The recommended procedure is:

1. A separate estimate of overall bias is made for each laboratory.

2. The mean and standard deviation of the laboratory biases is calculated. The mean is used as an estimate of method bias.

3. The standard deviation from 2) is divided by the square root of the number of laboratories to give an estimate of the standard error of the estimated method bias.

Notes

1. The laboratory biases to be obtained in step 1 are calculated by working out the average difference between results using the candidate method and the relevant reference values. However, in the case where each laboratory uses both the candidate and reference methods on each sample, there are advantages to be obtained by using the average difference between the laboratories results on the candidate and reference methods. Although the bias estimate for each laboratory is less precise, the average over laboratories is approximately the same and the resulting estimate of standard error does not need the adjustment to allow for uncertainties in the reference values described in Note 2 below.

2. If reference samples are used, the standard error obtained in 3) above should be adjusted to allow for uncertainty in the reference values. A conservative estimate would be to replace s_{bias}, the standard error of the bias calculated above by $\sqrt{s_{bias}^2 + s_{ref}^2}$, where s_{ref} is the average half-width of a supplied 95% confidence interval for the reference value. This value for s_{ref} assumes that all reference values were assigned using the same group of laboratories in the same trial. If the conditions under which the reference values were obtained are known to the required degree of detail, a more appropriate value may be used for s_{ref}.

3. Subsequent investigations of sensitivity may suggest that the bias may vary with analyte level. In such a case the bias calculated above is taken as applying at the average analyte level, calculated approximately by averaging all results obtained by the reference method. Calculation of confidence intervals for such a concentration-varying bias is discussed above.

SENSITIVITY

[Definition

Quotient of the change in the indication of a measuring system and the corresponding change in the value of the quantity being measured.{VIM}

Notes:

1. The sensitivity can depend on the value of the quantity being measured.
2. The change considered in the value of the quantity being measured must be large compared with the resolution of the measurement system.]

Estimation of sensitivity

In this context, linearity and sensitivity are taken as measuring the non-uniformity of bias over the range of analyte concentration concerned. According to the "black box" point of view being taken, the "indication of a measuring system" should be taken as the output estimate of concentration.

Sensitivity is estimated as a regression coefficient of estimated concentration on true concentration and should be close to unity. Linearity is tested as goodness of fit of the regression line.

The recommended estimation technique is similar to the one proposed above for overall bias:

1. A separate estimate of the regression coefficient is made for each laboratory.

2. The mean and standard deviation of these regression coefficients is calculated. The mean is used as an estimate of the overall method sensitivity.

3. The standard deviation from 2) is divided by the square root of the number of laboratories to give an estimate of the standard error of the estimated method sensitivity.

Often true concentration will be unavailable, and reference values may have to be used instead. This generally causes the method sensitivity to be under-estimated on average. To minimize this under-estimation, the uncertainties in the reference values need to be small compared to the range of concentrations over which the sensitivity is estimated. Even so, with increasing numbers of participating laboratories, it is possible that such a spurious under-estimate may be found statistically significant. The statistical significance of the estimate (as compared to unity) may thus need to be subordinated to consideration of its practical impact (as is probably appropriate in any case).

For comparing the range of reference values with their uncertainties, it is difficult to give criteria for smallness that are generally valid. The downward bias of the estimate of the regression coefficient β will be approximately $\beta\frac{\sigma_e^2}{\sigma_x^2}$, (provided $\frac{\sigma_e^2}{\sigma_x^2}$ is small,) where σ_e is the standard deviation of the errors in the reference values and σ_x the standard deviation of the reference values themselves, and the practical import of bias of this order should be considered. Sometimes it may be necessary to test the method over a range exceeding that of its proposed use. A fall back position, when uncertainty in the reference values precludes satisfactory estimation of sensitivity, is to resort to direct estimation of bias within various concentration ranges, using only samples which fall within these ranges.

LINEARITY

[Definition

The ability of a method of analysis, within a certain range, to provide an instrumental response or results linearly dependent on the concentration or amount of the analyte.]

General Comments on Linearity

In keeping with the black box approach, the "instrumental response" in the above definition is considered to be the output estimate of concentration.

The difficulties involved in testing for non-linearity are considerable. The root of the problem is that errors in the reference values are on average reproduced, with a change in sign, in the test results from the candidate method. This may not be important for testing results from a single laboratory, as the

repeatability errors involved probably swamp the errors in the consensus values, but as the number of laboratories increases, the repeatability errors are averaged out, while the errors in reference values are not.

With the minimum number of five samples suggested in the IUPAC guidelines, it is quite possible that patterns in the reference values (e.g. a predominantly convex shape) are present by chance, and would result in a finding of non-linearity, when the only true non-linearity present is in the reference values.

This problem also occurs in the estimation of sensitivity, but there the effect can be reduced by taking a sufficiently wide range of analyte concentrations. In the case of non-linearity this solution is not available.

The only solution may seem to be to considerably increase the number of samples used within the concentration range. Even then, a general test for non-linearity will tend to be failed as the number of laboratories increases, and the following method, which may seem at first sight to be naïve, is recommended.

This method is recommended because the denominator of the F-test proposed includes non-cubic variation from the errors in reference values as well as from the candidate method. Tests based on repeatability and reproducibility of the candidate method make no allowance for errors in consensus values.

Recommended procedure – linearity

1. For each laboratory and each sample, the duplicates for the candidate method should be averaged. For each sample, these averaged values are then plotted against the reference values, and a normal linear regression is performed. A computer program for linear regression that reports the associated analysis of variance table should be used.
2. Fit a straight line $Y = a + bX$, where Y is an average from the candidate method and X the reference value. Obtain the residual sum of squares and degrees of freedom (RSS_{Line} and RDF_{Line}).
3. Calculate $X2 = (X - \overline{X})^2$ and $X3 = (X - \overline{X})^3$. Add terms in X2 and X3 to the regression to fit a cubic curve. Obtain the residual sum of squares and degrees of freedom (RSS_{Curve} and RDF_{Curve}).
4. Combined test for quadratic and cubic curvature:

   Calculate the test statistic $\dfrac{(RSS_{Line} - RSS_{Curve}) \times RDF_{Curve}}{RSS_{Curve} \times (RDF_{Line} - RDF_{Curve})}$ and carry out an F- test. Statistically significant evidence of non-linearity (at the 5% level of significance) occurs when the test statistic exceeds the upper 5% point of Snedecor's F distribution, with $RDF_{Line} - RDF_{Curve}$ and RDF_{Curve} degrees of freedom. $RDF_{Line} - RDF_{Curve}$ will be 2.

Notes

1. The regression coefficient obtained in step 2 will be close to the estimate of sensitivity calculated by the recommended method. However, any standard error of the regression coefficient reported by the program will be different from the recommended one and should not be used in its place.
2. A program for stepwise regression could normally be used to perform this test. First fit the linear term, then the quadratic and cubic terms simultaneously, and see whether the added terms are judged statistically significant.

PRECISION (REPEATABILITY, REPRODUCIBILITY, REPRODUCIBILITY NET OF REPEATABILITY)

[Definition

The closeness of agreement between independent test/measurement results obtained under stipulated conditions.

Notes:

1. Precision depends only on the distribution of random errors and does not relate to the true value or to the specified value.

2. The measure of precision is usually expressed in terms of imprecision and computed as a standard deviation of the test results. Less precision is reflected by a larger standard deviation.

3. Quantitative measures of precision depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme conditions.

REFERENCES:

ISO Draft Standard 3534-2: Vocabulary and Symbols Part 2: Applied Statistics, ISO, Geneva, 2004]

Estimation

The criteria proposed in Annex B require the estimation of an upper 80% confidence interval for both the repeatability standard deviation and the standard deviation for reproducibility net of repeatability. For efficient estimation of the latter, it will normally be necessary to split the samples over several runs within each laboratory.

The statistical model involved is

$$X = x + e_{lab} + e_{run} + e_r$$

where

X is the measured concentration of a sample

x is the true concentration of the sample

e_{lab} is an error with mean zero and standard deviation σ_{lab}, which takes the same value for each measurement made by a particular laboratory

e_{run} is an error with mean zero and standard deviation σ_{run}, which takes the same value for each measurement made by a laboratory in a single run, but which varies from run to run within a laboratory

e_r is an error with mean zero and standard deviation σ_r, which takes a different value for every measurement made. This is identified with repeatability error.

The errors terms are assumed to be independent and normally distributed, and their standard deviations are considered to be constant.

Estimates and upper 80% confidence bounds are required for σ_r, the repeatability standard deviation, and

$$\sigma_L = \sqrt{\sigma_{lab}^2 + \sigma_{run}^2}$$

the standard deviation for reproducibility net of repeatability.

Various trial designs are possible, presenting various complications

1. Some designs, by the inclusion of repeatability and reproducibility duplicates, enable the direct estimation of repeatability and between run standard deviation from averages of squared differences between duplicates. However, particularly if precision parameters can be assumed to remain constant over the range of analyte levels presented, the trial may contain additional information on these precision parameters. The fitting of an appropriate random effects model will enable this additional information to be utilised, and would be expected to lead to more precise estimates of the standard deviations, at the expense of a more complicated analysis.

2. If the reference values are subject to negligible uncertainty compared to the repeatability of the method, and the precision parameters do not vary with analyte level, then if the appropriate reference value is subtracted from each result, the samples may be considered to lose their identity: all that is left is a collection of experimental errors, which may be analysed relatively simply by hierarchical analysis of variance.
3. If the reference values are unknown, or subject to appreciable uncertainty, then true analyte level must be included as a fixed effect in the analysis. This will involve the fitting of a mixed model (including both fixed and random effects.) It will also cause some designs to become non-orthogonal, requiring the use of more complicated estimation procedures such as REML. It will be desirable to take statistical advice on these matters, and also to check that appropriate computer software is available for the analysis of the results. It should be assumed that some outliers will need to be excluded from the analysis, and the proposed analysis method should allow for this.

In obtaining confidence limits for the standard deviations there are two steps:

1. obtaining estimates of σ_L^2 and σ_r^2 together with estimates of their sampling variances (the squares of their standard errors), and
2. using these to obtain confidence intervals for σ_L and σ_r.

1. Obtaining the estimates and estimates of their sampling variances

In estimating the required standard deviations and their confidence intervals, there are two possible starting points, depending on the type of analysis and software used.

a) Estimates of the variance components, σ_{lab}^2, σ_{run}^2 and σ_r^2, together with their covariance matrix. An estimate of σ_L^2 may then be obtained from $\sigma_L^2 = \sigma_{lab}^2 + \sigma_{run}^2$, and the sampling variance of the estimate is obtained from $\text{var}(\sigma_L^2) = \text{var}(\sigma_{lab}^2) + 2\,\text{cov}(\sigma_{lab}^2, \sigma_{run}^2) + \text{var}(\sigma_{run}^2)$

b) An analysis of variance table giving the following components

	Sum of Squares	Degrees of freedom	Mean Square
Between laboratories		df_A	MS_A
Between runs within laboratories		df_B	MS_B
Within runs		df_C	MS_C

together with formulae for the expected mean squares

$$E(MS_A) = \sigma_r^2 + K\sigma_{run}^2 + L\sigma_{lab}^2$$
$$E(MS_B) = \sigma_r^2 + M\sigma_{run}^2$$
$$E(MS_C) = \sigma_r^2$$

in which numerical values are given for the coefficients *K*, *L* and *M*. These may easily be solved to

express σ^2_{lab}, σ^2_{run} and σ^2_r, and thence σ^2_L, as linear functions of the expected mean squares. Estimates of these parameters may then be obtained by substituting the observed values of the mean squares for their expectations.

Estimates of the sampling variances of these estimates may then be obtained from the assumption that the mean squares are proportional to independent chi-squared variates. This gives

$$\text{var}(MS) = \frac{2 \times MS^2}{df}$$

for each mean square. Since the mean squares are independent, the sampling variances of the linear functions above may then be estimated from

$$\text{var}(\sum \alpha_i X_i) = \sum \alpha_i^2 \text{ var}(X_i),$$

a formula for finding the variance of a linear function of various independent components X_i from the variances of the separate components.

2 Obtaining the confidence intervals

Exact confidence intervals are available for σ_r based on the chi-squared distribution. In case b) above the appropriate number of degrees of freedom is df_C in the table. In case a) the degrees of freedom can be calculated as

$$df = \frac{2 \times \sigma_r^2}{\text{var}(\sigma_r^2)}$$

Although the estimate of σ^2_L will not be proportional to a chi-squared variate, a chi-squared approximation (Satterthwaite's Approximation) is often used. The number of degrees of freedom is given by the same procedure, namely

$$df = \frac{2 \times \sigma_L^2}{\text{var}(\sigma_L^2)}$$

where σ^2_L and $\text{var}(\sigma^2_L)$ are calculated as described above. In this case *df* will not normally be an integer.

An upper 80% confidence limit for σ_r is then calculated as

$$\sigma_r \sqrt{\frac{df}{\chi^2_{df,0.20}}}$$

where $\chi^2_{df,0.20}$ is the lower 20% point of the chi-squared distribution with *df* degrees of freedom.

A similar procedure is then used to obtain an upper 80% limit for σ_L.

Examples of the calculations involved are given in Annex C.

Note

Some statistical packages have options (which may be the default) to constrain the estimates of the components σ^2_{run} and σ^2_{lab} to be both positive. While this may be logically satisfying, the required

end product of the investigation is not σ^2_{run} and σ^2_{lab}, but $\sigma^2_{lab} + \sigma^2_{run}$, and the estimate of this may be biased by introduction of such constraints. They have also been observed to sometimes lead to numerical instability in the estimates. Thus such options should be avoided if possible. Biased estimates of $\sigma^2_{lab} + \sigma^2_{run}$ will also be obtained if negative estimates of σ^2_{lab} are replaced by zero.

LIMIT OF DETECTION

[Definition

The concentration of an analyte corresponding to the lowest measurement or measurement signal which with a certain statistical confidence may be interpreted as indicating that the analyte is present in the sample, but not necessarily allowing quantification.]

Estimation

The limit of detection is determined by $LoD = m + 3s$, where m is the mean and s the standard deviation of estimates of concentration from the testing of blank samples under reproducibility conditions.

This estimation corresponds, at least nominally, to a 0.1% chance of declaring the analyte present when the result truly arises from a blank sample.

Notes

1. The limit of detection is defined by the response to blank samples. The response to non-blank samples is only relevant if needed to determine the calculations involved in stating this response in terms of concentration, for example by means of a calibration curve.
2. Frequently a definition is given from which the term m is omitted. It is not easy to account for this, and it is thought that such a definition must have originated in rather a specialized context.

LIMIT OF QUANTIFICATION

[Definition

The lowest concentration of analyte in a sample which can be measured with a certain statistical level of confidence.]

General Comments on Limit of Quantification

The meaning of this definition is not altogether clear, but it is assumed that what is required is a concentration above which the relative reproducibility is reasonably small. To estimate such a limit a definition of "reasonably small" is required. This will vary with the context in which the method is applied.

For some analytes and methods, the relative reproducibility could be approximately constant, and such a requirement may never be fulfilled. In other cases, the reproducibility itself may be approximately constant, and if reproducibility limit of less than k% of concentration is required, the limit of quantification will be simply $\frac{100 \times R}{k}$, where R is the reproducibility limit. In yet other cases, estimates of relative reproducibility may have to be made for each sample and plotted against concentration to estimate the LoQ by eye (the variation seems unlikely to be linear).

But in any case, adequate reporting of the reproducibility (as a function of concentration if there is substantial evidence that it varies with concentration) will enable the LoQ to be estimated as the need arises, without any need to prejudge a value of k to be applied in all contexts.

Before the reporting of a limit of quantification could be considered useful, it would be necessary to agree on an unambiguous definition that summarises some meaningful characteristic of method performance in a way that can be usefully applied.

APPLICABILITY

[Definition

The analytes, matrices and concentrations for which a method of analysis may be used satisfactorily to determine compliance with a Codex Standard.]

Notes

1. Applicability should also include a demonstration of the ruggedness of the method.

2. Although restrictions on the group of laboratories to which the performance parameters of the method can be applied are not mentioned in the definition, some such restriction may be implied by the trial design, particularly if the validation trial involves fewer than the normal minimum.

Confirmation

The analytes, matrices and range of concentrations for which a method may be used are normally established during the development and initial validation of a test method.

Before applying the method to a different matrix, or to measure a concentration outside the range covered by the existing method, it is first necessary to confirm that the method performs satisfactorily on the new matrix. At a minimum this confirmation should verify that the bias and repeatability of the method for the new matrix and in the new range are consistent with the established characteristics for the method.

The assessment of ruggedness is discussed separately below.

RECOVERY

[Definition

The proportion of the amount of analyte present or added to the test material which is extracted and presented for measurement.]

Estimation

Recovery is determined by adding known quantities of the analyte, for example as a known volume of a solution of known concentration. The concentration of the added sample will be known either from the use of a pure compound or, say, from the use of a reference sample.

The following example serves to illustrate the principle for the estimation of recovery.

Firstly, the original sample is analysed several times and the results averaged to provide an estimate of the actual concentration in that sample, the value *O*.

Second, a portion of the original sample is taken and spiked with an accurately measured amount of the analyte, to increase the concentration by a known amount *S*. This second sample is analysed several times to provide an estimate *F* of the concentration.

The percentage recovery for the method is then calculated using:

$$\%\ \text{Recovery} = 100 \times \frac{F - O}{S}$$

where

F is the average result from testing the final sample

O is the average result from testing the original unspiked sample, and

S is the known increase in concentration introduced by spiking.

Notes:

1. Where the recovery is not equal to 100% there is a bias of the results relative to the true values. As assessment must be made to determine whether this bias is of practical significance and any remedial actions made, such as the application of a recovery correction.
2. As recovery rates may vary between laboratories the estimation of recovery should be undertaken across several laboratories. In this case the reported recovery rate for a method would be an average across laboratories.
3. The estimation technique overcomes problems caused by the presence of an unknown level of the analyte in the samples tested.
4. The inherent problem with this technique is that the form of the analyte introduced in the spiked sample may not be the same as that in the original sample, so that recovery of the spiked material might not provide a reliable indication of the recovery of the analyte in the original sample. For this reason a variety of different samples, each with a different matrix, should be used for spiking.

RUGGEDNESS (ROBUSTNESS)

[Definition

The ability of a measurement process to resist, in terms of the effect on test results produced, deviations made from the experimental conditions described in the method.]

Estimation

The investigation of robustness is undertaken by deliberately varying the conditions under which the method is carried out and determining the effect on the test results produced.

Fractional factorial designs, as suggested by Youden, are one method commonly used to assess ruggedness. In this approach, the conditions under which the method is carried out, for example changes in the instrument, the operator, the brand or concentrations of reagents used, or the temperatures or times for heating, are deliberately varied about the levels specified in the method according to a specified statistical experimental design. The analysis of the data will show the effects of changes in each of the conditions on results generated by the method, and their interactions.

Notes

1. It is recommended that a statistician be consulted for advice about suitable experimental designs, the analysis of the data from the trial and interpretation of the results.
2. Some guidance may be obtained from statistical texts, two of which are referenced.
3. The method may have to be reviewed and even revised in the light of the outcomes from a ruggedness study. Alternatively the outcome of a ruggedness study may lead to restrictions being placed on the use of a method, for example that it is not suitable for use in a certain concentration range.

References:

- Cochran W.G. and Cox G.M. Experimental Designs (2 ed) John Wiley & Sons 1957
- Box G.E.P., Hunter W.G, Hunter J.S. Statistics for Experimenters John Wiley & Sons 1978

SELECTIVITY

[Definition

Capability of a measuring system, using a specified measurement procedure to provide measurement results for two or more quantities of the same kind involving different components in a system undergoing measurement, without interference from each other or from the quantities of the system. {VIM}]

Estimation

The selectivity of a method is usually investigated by studying its ability to measure the analyte of interest in test portions to which specific interferences have been deliberately introduced.

Immediately, the techniques proposed to investigate recovery might be applied to investigate selectivity. A ruggedness-type study might be also used to measure the effect of specific interferences, or the interaction of different interferences, on the analyte of interest.

EXPLANATORY NOTES FOR ANNEX A: ESTIMATION OF CHARACTERISTICS

MEANING AND USE OF METHOD PERFORMANCE PARAMETERS

The parameters to be estimated are not those quantifying the accuracy of which the method is intrinsically capable, but those quantifying the accuracy to be expected when laboratories use the method under normal operating conditions. They are necessary so that appropriate tolerances or safety margins can be used in a regulatory environment.

Acceptability of a candidate method will then be a question of whether the tolerances and safety margins appropriate to the currently used method, if there is one, continue to be appropriate when the candidate method is used or whether they will need widening to an unacceptable extent. The acceptability of a method where no other method is currently in use will depend on whether the necessary tolerances are acceptable in themselves in a given regulatory environment. If not, the method may have to be used none-the-less for lack of an alternative, and unacceptability will mean that development of an alternative method is necessary.

Normally the sample of laboratories involved in a validation trail will be small, in the statistical sense, and the time over which the trial is conducted will be limited. It would be expected that an attempt would be made to check that laboratories continue to perform within the limits suggested by the validation trial, on the basis of which tolerances and safety margins are set, and to review these tolerances in the light of further evidence if it should transpire that the performance parameters need adjustment. Whether such monitoring is or could be put in place may itself be a relevant consideration in deciding whether a method is appropriate.

DEFINING METHODS

Although a method may be a defining method, this does not imply that laboratory and run biases cannot exist within the method, and in fact there may exist other unbiased methods with better precision. Thus, although defining (Type 1) methods are excluded from the scope of these guidelines, it is worth noting that it is conceivable that such a method could reasonably be deemed unacceptable as a replacement for a more precise rival in some circumstances.

SCOPE OF VALIDITY OF AN ASSESSMENT OF METHOD PERFORMANCE

The scope of the assessment is determined by the range of experimental conditions under which the method has been tested. As well as limits to the range of concentrations and matrices for which the performance of the method may be considered to have been established, there may be other limits to be taken into account.

For example, a trial performed by a single laboratory by a single operator in a single run would establish the performance only on that run. If the trial is carried out in a single laboratory by a single operator over several runs that can be considered representative of normal operating conditions, then the assessment would in general extend at most to that operator in that laboratory under normal operating conditions.

If the validation is performed by a single laboratory using several appropriately certified operators under normal operating conditions over a considerable period of time, then the method performance could be considered as established for that laboratory under those conditions.

Only if the method is tested using a reasonable number of representative laboratories can the performance of the method itself be assessed directly. Extrapolation of results based on more limited data may sometimes be necessary as a stopgap measure.

It should be noted that a particular laboratory or group of laboratories could, in principle, claim validity for a method that has been found unsuitable in general, by demonstrating superior performance in the use of that method. This may be acceptable in some circumstances. However, it should be noted that stability over time would be much more of an issue in such a case. In an inter-laboratory trial participating laboratories may each vary in performance over time, while the “group behaviour” remains stable. The use of a substantial number of laboratories in a trial results in an

automatic randomization over various potential sources of bias such as calibration errors in scales or thermostats, variation in strengths of stock solutions, between-batch variation in microscope slide markings and so on. It may be hard to ensure representative variation in these factors over a relatively short time frame within a single laboratory, whereas over a longer time frame, even within a single laboratory, they may cause significant variation.

NEED FOR INTER-LABORATORY ESTIMATES OF BIAS

The fact that the reproducibility of a method is normally considerably higher than its repeatability demonstrates that the existence of significant laboratory biases must be taken as the norm. Results from a single laboratory over a short period of time have to be assumed subject to a bias, relative to the true analyte level, with standard deviation $\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$. Typically σ_L will be about 85% of σ_R, and so laboratory biases may reasonably be expected to fall within a range of +/- $1.7\sigma_R$ of the average laboratory bias for the method (the method bias.)

In some contexts, measurement errors of the order of $1.7\sigma_R$ may not be significant. These will be largely the contexts in which measurement error as a whole can reasonably be ignored. In any other context, results from a single laboratory will not be adequate either to confirm the absence of a method bias or to estimate it if it is present: In either case, the result could be out by $1.7\sigma_R$.

Thus it is highly desirable to incorporate the estimation of bias into a procedure involving a reasonable number of laboratories. It is therefore assumed that this part of the validation is included in the collaborative trial in which reproducibility of the method is estimated. The exceptions would be cases in which it was confidently expected that the method reproducibility would prove negligible in all uses envisaged for the method.

This implies either the use of reference samples in the collaborative trial, or the estimation of reference values within the trial itself. Reference samples could, depending on chemistry and practical considerations, be either artificially generated samples of known analyte concentration, or samples whose values had been estimated in a collaborative trial using the reference method. In any case, uncertainties regarding their true analyte concentrations need to be considered, if only to demonstrate their negligibility.

EFFECT OF UNCERTAINTY IN REFERENCE VALUES

- on estimation of overall bias

Assessment of overall bias is essentially a matter of comparing the average result using the candidate method with an appropriate average of reference values. A supplier of reference samples will normally have a pool of laboratories on which he can call, and will normally submit more than one sample for analysis on any given occasion. As has been discussed above, laboratory bias is a substantial component of the measurement error. The same laboratory biases will attach to each sample submitted to the pool of laboratories on this occasion, and will not be reduced by averaging over samples. Even for samples submitted to the same pool on different occasions, there will be a major component of bias in common. Thus reference values for different reference samples cannot automatically be assumed to be statistically independent. Averaging over a large number of reference samples in this situation will not cause the measurement error associated with the average to tend to zero, but rather to tend to a linear combination of a relatively small number of laboratory biases. The number of laboratories involved in the determination of reference samples may well be comparable to or even less than the number of laboratories involved in the method validation. In the latter case, uncertainty surrounding the reference values may well be the dominant cause of uncertainty in the estimate of method bias.

The history of the reference samples will probably not be known in sufficient detail for accurate calculation of this uncertainty. While reference samples may come with an indication of the precision of the reference vales, and an indication of the number of laboratories involved in their estimation, it would not normally be clear whether the 10 laboratories on which the reference value for sample A is based are the same as the 10 on which that for sample B is based, and if so, whether they analysed

samples A and B on the same or different runs. Troublesome requirements for details such as these are unfortunately necessary for a correct statistical treatment. If the details are unknown it is necessary to resort to conservative estimates.The total uncertainty attaching to a single measurement after correction for bias standard deviation

will be $\sqrt{\sigma_R^2 + \sigma_{bias}^2}$, where σ_{bias} is the standard deviation measuring the uncertainty of the bias. Sometimes σ_{bias} will make a negligible contribution to this expression, and thus may appear at first sight unimportant. However, the criteria proposed in Annex B involve a separate tolerance for bias and σ_{bias} may well have a significant impact on this tolerance. The matter is further discussed in the discussion paper to Annex B.

- in estimating sensitivity

In general it may be shown that measurement errors in the independent variable, here reference values of other estimates of concentration under the standard method, cause the regression coefficient to be underestimated on average.

The problem being dealt with is one of structural relationship, where both the dependent and independent variables are subject to measurement error and the parameter of interest is the response to the true, and not the measured value, of the independent variable. This may be illustrated by an extreme example. Suppose all the concentrations are in fact equal, but due to substantial measurement errors in the reference method appear to vary significantly. Failure of the candidate method to respond to this fictitious variation is then entirely proper, and we should not expect a regression coefficient of unity.

At least in simple cases the expectation of the estimated regression coefficient, as computed by normal least squares, is approximately $\beta\left(1-\frac{\sigma_e^2}{\sigma_x^2}\right)$, provided the second term in brackets is small. Here σ_e and σ_x are standard deviations measuring the uncertainty attached to each true value as measured by the reference method, and the dispersion of these measured values, respectively. β is the true sensitivity. So that the severe complications involved in the estimation and testing of structural relationships may be avoided, it will be necessary that σ_x be large compared to σ_e. Unfortunately this brings its own problems, as the larger σ_x becomes, the more precisely the regression coefficient is estimated, leading to a chance that even a small bias may be found statistically significant.

The standard error of estimation of the regression coefficient (as estimated by a single laboratory in a single run) is given by $\sigma_{\hat{\beta}} = \frac{1}{\sqrt{n}} \times \frac{\sigma_r}{\sigma_x}$ where σ_r is the repeatability of the candidate method. To avoid spurious findings of statistical significance, the bias must be small compared to this, that is, we must have

$$\frac{1}{\sqrt{n}}\frac{\sigma_r}{\sigma_x} >> \frac{\sigma_e^2}{\sigma_x^2}$$

This reduces to

$$\frac{\sigma_x^2}{\sigma_e^2} >> n\frac{\sigma_e^2}{\sigma_r^2}$$

Depending on the context, this may or may not represent a significant constraint. For example, if a laboratory analyses 10 duplicate pairs of samples using both the candidate and reference methods, and the repeatability of the two methods are the same, *n* will be 20 and σ_e^2 will be about half σ_r^2. If we require the bias of the estimate of sensitivity to be less than 20% of the standard error, we then have $\sigma_x > 16\sigma_e$, that is $\sigma_x > 8\sigma_r$. Assuming that reference values are uniformly spread over the range of concentrations used, this translates into a condition that the range should exceed $28\sigma_r$.

Where the coefficient is estimated as an average over several laboratories the standard error with which the coefficient is estimated may be further reduced, particularly in the absence of true inter-laboratory variation in sensitivity, and the risk of spurious findings of statistical significance is increased. Thus stress is laid in Annex A on designing the trial so that the bias can be expected to be unimportant, rather than on tests of statistical significance.

It should be noted that the sensitivity is the slope of a line, and is thus unaffected by a uniform increase or decrease in reference values: only differences between reference values is relevant. If all reference values used have been assigned by the same group of laboratories, differences between the reference samples will normally be considerably more precise than their individual uncertainties may suggest. In extreme cases, the differences will be affected only by repeatability error, whereas the uncertainties supplied will be based on reproducibility type error. If enough is known about the means by which the reference values were estimated, and in particular if they were estimated as part of the validation trial itself, account may be taken of this in judging the potential for bias in the resulting estimates of sensitivity. An example (Example 3) is given in Annex C.

ANNEX B: CONDITIONS FOR ACCEPTANCE OF METHODS

INTRODUCTION

This annex describes conditions under which a candidate method may be accepted as a general replacement for a standard method in judging product compliance after a method validation exercise. After acceptance, the candidate method could then be used in any context (involving testing for product compliance) in which the standard method is authorised and the applicability of the candidate method has been established, without the need for additional tolerances. The conditions have been formulated to avoid increasing an assumed producer's risk of 5% to more than 7.5%.

It must be emphasised that failure to meet the conditions does not exclude a method from use in particular and specified circumstances. The validation exercise will provide information that can be used to assess the risks attached to the use of the method in such circumstances and thence judge its fitness for purpose. However, it is envisaged that only exceptional circumstances would justify objections to the use of a candidate method that does meet the conditions.

Note

In some cases there may be a multitude of "standard methods", or indeed no other method at all. While the guidelines do not deal explicitly with these situations, it is clear that judging a method acceptable implies comparison with some standard.

Where several methods exist, it may be appropriate to include between-method variation as part of the reproducibility, running a validation trial over laboratories using different methods. Where no method exists, standards would have to be postulated in advance.

RATIONALE

The replacement of one method by another will always involve an element of risk either to producer or consumer or both. The conditions on precision parameters are based on the premise that product that has a 5% probability of failure in a compliance test, using the standard method, should not face a potential probability of failure of more than 7.5% if the candidate method is used instead. Without further specification of the context, this cannot be translated directly into conditions on method performance parameters. However, in the attached discussion paper a variety of compliance tests are considered, and a conclusion is reached that increases above 14% in the standard deviations for repeatability or reproducibility net of repeatability seems likely to breach the requirement in respect of at least some of them, unless compensating tolerances are introduced into the relevant compliance testing procedures.

The estimates of the standard deviations obtained from a validation trial are subject to uncertainty that may be considerable, and it is desirable to control the risk that a candidate method will be accepted when in truth it does not meet the requirements. This risk has been set at 20%. While this may seem large, a substantial reduction in this risk seems to lead to increases in size of the validation trial that are probably not practical. The matter is discussed in the attached discussion paper.

In addition to controls on the standard deviations, adjustments for method bias are necessary.

CONDITIONS

Trial design

To plan and carry out a validation exercise is a major undertaking involving scientific and statistical considerations that will vary considerably according to the analyte and method being considered. No prescription is therefore made on the type of investigation or the methods of estimation that should be used, other than that they should yield scientifically and statistically valid estimates or assessments of the performance characteristics listed below. It is however thought unlikely that satisfactory estimates of bias and precision will be obtained unless a substantial inter-laboratory trial is included. Possible

estimation methods for performance characteristics are discussed in Annex A, but these should be considered as suggestions rather than prescriptions.

Bias

Upper and lower one-sided 95% confidence limits for method bias (or relative method bias, if this is more appropriate) at various concentrations within the range should be given, for the purpose of calculating appropriate adjustments when product is tested against a specification limit. (Note that together these limits will form a 95% confidence interval.)

Sensitivity

No conditions are placed on sensitivity, partly due to the difficulty of estimating it in certain circumstances. Estimates of sensitivity when these can be made will normally be incorporated into the confidence intervals for bias.

Linearity

Statistically significant evidence of non-linearity should be reported, and the confidence bounds for method bias must be calculated in a way that allows for the non-linearity. This would normally be done by calculating independent estimates of bias at various parts of the range.

Precision

An upper 80% confidence bound should be calculated for the repeatability standard deviation. This should not exceed the accepted value for the repeatability standard deviation of the standard method by more than 14%.

An upper 80% confidence bound should be calculated for the standard deviation for repeatability net of reproducibility ($\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$). This should not exceed the accepted value for the corresponding standard deviation of the standard method by more than 14%.

An upper 80% bound for the reproducibility standard deviation should also be given.

Failure to meet these requirements would require a tolerance to allow for the possible additional measurement uncertainty. An example of the calculation of such a tolerance is given in Annex C.

Limit of detection[, limit of quantification]

Should be reported.

Applicability

The range of matrices and concentrations for which the method has been tested and found appropriate should be given. Known matrices for which the method is unsatisfactory should be reported.

Recovery

Recovery rate, if not explicitly corrected for as part of the methods will form an element of the method bias, and accordingly no additional conditions are required. Anomalous recovery rates for particular matrices should be reported, and the matrices concerned excluded from the range of applicability of the method.

Ruggedness

The dependence of results on variations in experimental conditions should be investigated, and attention should be drawn to conditions that need to be particularly carefully controlled.

Selectivity (interference effects etc.)

Any interference effects should be reported.

NOTE

Failure of a method to meet the precision criteria means that an additional tolerance needs to be applied in compliance testing when the method is used, to allow for the possible additional

measurement uncertainty. This would be applied on top of the adjustment for method bias. There is a possibility that increased measurement uncertainty could be compensated for (from the producer's point of view) by, for example, a negative upper 95% bound for bias when testing against an upper limit. This could render the method fit for purpose in some contexts. However, it is not currently proposed that such a method should be considered "generally acceptable," as the method would on the face of it appear likely to be inferior, with a statistically significant bias and at least the possibility of substantially poorer precision than the standard method.

EXPLANATORY NOTES FOR ANNEX B: CONDITIONS FOR ACCEPTANCE OF METHODS

INTRODUCTION

These notes address the conditions under which a candidate method may be accepted as a general replacement for a standard method in judging product compliance, after a method validation exercise involving one or more laboratories. After acceptance, the candidate method could then be used in any context (involving testing for product compliance) in which the standard method is authorised.

There is wide range of possible situations to be covered, ranging from cases of dangerous contamination where a producer has no business to be anywhere near the compliance limit to cases where a producer needs to work hard to keep within a narrow specification, with no serious health implications if he moves outside them. The compliance testing methods involved range from simple cases where a single sample is taken and checked against a compliance limit, to complicated situations where sample means and standard deviations are tested both alone and in combination. The product being assessed may be consistent or very variable, depending not only on the product type but also on the individual producer or even the individual lot.

For a method to be acceptable in all such circumstances seems rather a tall order, and the criteria for acceptance would be expected to be rather onerous. This indeed seems to be the case. It must be emphasised that a method that is not found generally acceptable may be more than adequate in particular circumstances. However, more detailed consideration would have to be given to the particular circumstances involved.

SUMMARY

The criteria suggested by the discussion below are, that for a candidate method to be considered acceptable as a replacement for a standard method, the following criteria should be satisfied.

a) One-sided upper and lower 95% confidence intervals should be provided for the method bias, to adjust any upper or lower cut-off value used in compliance testing. (Together these would constitute a 90% confidence interval.)

b) That a one-sided upper 80% confidence limit for the repeatability standard deviation should not exceed the repeatability standard deviation for the standard method by more than 14%.

c) That a one-sided upper 80% confidence limit for $\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$ should not exceed σ_L for the standard method by more than 14%.

These criteria are formulated to control the likelihood of large increase in producer's risk when the method is applied. Essentially a 20% risk is accepted of accepting a method with a potential to increase the producer's risk from 5% to 7.5%.

While this may seem rather lenient the requirements are believed to be rather difficult to meet unless a candidate method is in fact significantly better than the standard method. The chance of acceptance of a method that is comparable to the standard, even with a comprehensive validation trial, are assessed as better than even but by no means good.

In fact there seems rather little room to manoeuvre, with any significant improvement in one direction being matched by a significant deterioration in the other.

Most of the difficulty probably comes from the generality required of the criteria. As noted in the introduction, a method failing to meet the criteria may still turn out to be acceptable when specific consideration is given to the analyte, product and compliance testing procedures concerned. There is also a possibility, which has not been explored, that an initially unacceptable method may be rendered acceptable by the specification of additional tolerances for use in compliance testing.

DISCUSSION

TERMINOLOGY: PRODUCER'S AND CONSUMER'S RISK

- The producer supplies the product being tested.
- The consumer carries out a compliance test on the supplied product to determine whether it will be accepted or rejected.
- The producer's risk is the probability that a lot that would be considered acceptable if its true composition were known will be rejected by the compliance test. This varies with the true composition of the lot. In determining a sampling plan one consideration is that this risk should not exceed a certain level (usually 5%) for lots of certain true composition, for example with a given mean and standard deviation for the analyte being tested. We shall refer to this specified composition, for which the producer's risk is set, as the "acceptable quality level" by analogy with the corresponding concept in "sampling by attribute."
- The consumer's risk is the probability that product of a given composition that is outside the compliance limits will be accepted.

Often a compliance test is decided on without explicit statement of these risks. In such a case we have taken the approach that an acceptable quality level has been implicitly determined by the compliance test itself, such that the producer's risk is 5%. A similar approach is possible to consumer's risk. However, the safety of the consumer often lies less in the fact that the compliance test has been passed in respect of any particular lot of product, than in the adverse consequences of frequent failure to the producer. These consequences force him to exert considerable effort to supply only product that will fall within the acceptable quality range.

USE OF TOLERANCES FOR MEASUREMENT ERROR

Normally decisions will be made on the acceptability of product by taking a sample of product and calculating a test statistic *S*. This is then compared to a cut-off level S_U. We will assume that this is an upper limit. Various sampling schemes listed in Codex documents and elsewhere deal with appropriate test statistics and cut-off levels in the absence of measurement error. In the presence of measurement error, particularly when, as is usually the case, this error has more than one component (for example repeatability error and between-laboratory error) the problem becomes considerably more complicated, and in at least some cases (for example, inspection by variables schemes) has not been solved satisfactorily.

An approach sometimes used is to use the same sampling scheme, using measured values instead of true values in calculating the test statistic. The cut-off level S_U is then increased to allow for measurement error, say to a value S'_U. The difference between S_U and S'_U can then be considered as a tolerance to allow for measurement error. Some product testing procedures may explicitly give such tolerances, others may ignore the presence of measurement error, and not specify the use of such tolerances. However, in either case consideration of an appropriate tolerance is relevant. If no tolerance is allowed for in the test procedure, then a producer must work to a de facto limit lower than the specification limit in order to control his risk. (This de facto limit corresponds to the "acceptable quality level" of acceptance sampling).Unfortunately, exact calculations of the appropriate value of S'_U necessary to set the producer's risk at a suitable level will normally involve process parameters as well as the parameters describing the distribution of the measurement error, if indeed such a value can be calculated at all.

The cut-off level could also in principle be reduced to a value designed to control the consumer's risk, with an increased risk to the producer. This would involve calculating S'_U by subtracting, rather than adding, a tolerance. This seems undesirable, as

a) In rejecting product from a reputable source as non-compliant, the onus is normally considered to lie on the consumer to prove that the product is non-compliant, rather than on the producer to prove that it is compliant, and

b) as sampling schemes, sample sizes and analytical test methods are normally under the control of the consumer, inadequacies in these should be paid for in terms of consumer's, rather than producer's, risk.

GENERAL APPROACH

The approach used has been to estimate appropriate tolerances for the various compliance testing procedures and examine the effect on producer's risk when a tolerance appropriate to the standard method is used with the candidate method. If the producer's risk rises to unacceptable levels the candidate method is considered unacceptable. This implies a need to calculate a more appropriate tolerance using the relevant error parameters for the candidate method: in effect, to use an additional tolerance when the method is applied.

Although the exact calculation of an appropriate tolerance may depend on process parameters, an upper limit can sometimes be found that does not depend on these parameters, and enables the producer's risk to be controlled.

Consider the following simple situation:

A normally distributed test statistic S is being tested against a cut-off S_U. At the acceptable quality level S had mean μ_S and standard deviation σ_S. Then for a producer's risk α, S_U will be set at

$$S_U = \mu_S + k_a \sigma_S$$

where k_α is the appropriate percentage point of the normal distribution. Now suppose that to S is added an independent measurement error of mean zero and standard deviation σ_M.

To maintain the producer's risk at α, the new cut-off S'_U should be

$$S'_U = \mu_S + k_\alpha \sqrt{\sigma_S^2 + \sigma_M^2}$$

giving an appropriate tolerance of

$$T = S'_U - S_U = k_a \left(\sqrt{\sigma_S^2 + \sigma_M^2} - \sigma_S \right)$$

This depends on σ_S. In the absence of information about σ_S, the best we can do is to give an upper limit, which is

$$T = k_\alpha \sigma_M$$

This upper limit can be interpreted as an upper confidence bound on the measurement error, and can in fact be obtained by formalising an argument that since M is unlikely to exceed the tolerance, S is likely to exceed S_U.

The upper bound will be very conservative if σ_S is comparable to or exceeds σ_M, and good if σ_S is substantially less than σ_M. To test acceptability of methods on the basis of their effect on this upper bound may at first sight seem over-stringent. However, it must be borne in mind that **ON SOME OCCASIONS WHEN THE METHOD IS USED, S MAY BE A MEAN**, either explicit or implicit, as when a several samples are combined into one composite sample. In such a case σ_S will tend to zero with increasing sample size, whereas σ_M, which contains a between-laboratories component of error, will not. Thus in considering acceptability for general purposes, the use of the conservative upper bound above is necessary.

EXAMPLES OF TOLERANCES FOR VARIOUS TESTS

In the following examples, no allowance has been made for method bias, or the uncertainty of its estimation. A discussion of the treatment of bias is given later.

In the examples, below, the following notation is used consistently.

- σ_r is the average repeatability standard deviation, σ_R the reproducibility standard deviation and σ_L is the standard deviation for reproducibility net of repeatability error, given by $\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2}$.
- k_α is the upper 100α percentile of the normal distribution giving producer's risk α. Usually k_α will be 1.645, corresponding to $\alpha = 0.05$ and giving a one-sided 95% confidence interval.
- n is the number of independent samples being tested and *d* is the number of duplicates of each sample, which are assumed to be averaged to give a single result for each sample before further computations take place. If there are no duplicates, *d*=1.

Other notation will be dealt with as it arises.

Test of a single (possibly composite) sample against a cut-off

$$T = k_\alpha \sqrt{\sigma_L^2 + \frac{\sigma_r^2}{d}}$$

Test of the mean of several samples against a cut-off

$$T = k_\alpha \sqrt{\sigma_L^2 + \frac{\sigma_r^2}{nd}}$$

The original cut-off S_U will often be based on an assumed value of the lot standard deviation, and is sometimes combined with a test that this assumed value is not inappropriate. It should be noted that the presence of measurement error will cause the standard deviation of a sample of measured values to exceed the true lot standard deviation on average. See example 0.

Inspection by Variables

The test is to reject when $m + qs > S_U$. Here m is the sample mean, s the sample standard deviation and q a coefficient chosen to give an appropriate rejection probability. The suggested tolerance is

$$T = k_\alpha \sqrt{\sigma_L^2 + \left(1 + \frac{q^2}{2}\right)\frac{\sigma_r^2}{nd}}$$

which uses a normal approximation to the distribution of the sample variance. It may be that a better estimate is available.

Test of several samples with failure if any one exceeds a cut-off

$$T = k_{\frac{\alpha}{n}} \sqrt{\sigma_L^2 + \frac{\sigma_r^2}{d}}$$

This estimate is conservative. It is based on using Bonferroni's inequality to find a conservative upper confidence bound for the largest of the *n* measurement errors.

Test of the square of the sample standard deviation against a cut-off based on the chi-squared distribution

The test being referred to is the test $s^2 > \frac{\sigma^2 \chi^2_{n-1;\beta}}{n-1}$ to test whether the observed sample standard deviation *s* exceeds an assumed value σ of the lot standard deviation. $\chi^2_{n-1;\beta}$ is the upper 100β percentage point of the chi-squared distribution on n-1 degrees of freedom, with β the size of the

significant test being used. Allowance for measurement error can be made exactly, assuming normality and that all samples involved are analysed in a single run (as they should be wherever possible.) The confidence interval approach is not needed. Denoting the right hand side of the inequality above by S_U the tolerance is

$$S'_U - S_U = \frac{\sigma_r^2}{d} \cdot \frac{\chi^2_{n-1;\beta}}{n-1}$$

Note that the tolerance is applied to a cut-off for the square of the standard deviation, not for the standard deviation itself.

CONSIDERATION OF THE TESTS IN 0

Whether a method is suitable for general use will in general depend on whether the tolerances required under the candidate method are significantly larger than those obtained using the standard method. Taking into account the requirement that process parameters should not appear in the criteria for method acceptance, and considering the above to be a reasonably representative collection of sampling procedures which may be used, it is proposed that a candidate method should be considered a generally suitable replacement for a specified method provided that the necessary tolerances listed above are not increased so as to cause a substantial increase in producer's risk when tolerances suitable to the standard method are used instead.

TESTS 0 TO 0

The expressions in 0 to 0 above are very similar, involving a single contribution from σ_L^2 together with various multiples, decreasing with sample size, of σ_r^2 . To see what sort of increase might be permissible, suppose that the tolerance itself is held fixed, but the square root factor has increased. Then the effective value of k_α being used is reduced in proportion, with a consequent increase in α , the possible producer's risk (possible, rather than actual, because we are dealing with an upper bound, although in some circumstances a realistic one.) If an increase in possible producer's risk from 0.05 to 0.075 is considered acceptable, k_α can go from 1.645 to 1.440, and the square root can increase by 14%. The increase can be limited to this amount by specifying that neither σ_L nor σ_r should increase by more than 14%. An increase in possible consumer's risk from 0.05 to 0.10 would probably be considered unacceptable and this would be generated by increases in σ_L and σ_r of 28%.

TEST 0

The tolerance given in 0 cannot be considered altogether satisfactory, because it is not clear how good the upper bound given by Bonferroni's inequality is. The errors will usually be very strongly correlated, having a major component, the run bias, in common. As it stands, the tolerance given is fairly sensitive to changes in the square root factor. An increase of 14% would change an α of 0.05 into one of 0.087 when n=2, and the situation rapidly becomes worse for increasing n. To limit α to 0.075 for two samples, we need an increase of no more than 10%, for five samples, no more than 7%, for ten samples no more than 6%. A more exact analysis may be possible that shows the situation to be not as bad as it seems. Alternatively, it may well be that this compliance test is not really appropriate when measurement error is a significant factor.

THE CHI-SQUARED TEST

The test for the sample standard deviation in 0 also needs special consideration and turns out to be very sensitive to changes in repeatability variation. This is not surprising, as it is designed as the most powerful test for detecting a change in standard deviation, and a change in repeatability error will directly cause such a change, causing measured values to be more dispersed that true values. Some product variation must be allowed for, namely the assumed lot standard deviation σ with which consistency is being tested. It is also assumed that the repeatability error of the standard method is taken into account. This gives the test

$$s^2 < \left(\sigma^2 + \frac{\sigma_r^2}{d} \right) . \frac{\chi^2_{n-1;\alpha}}{n-1}$$

and the question is what change in α is required to compensate for a given increase in σ_r ? The answer depends rather critically on the relative sizes of the two terms in brackets, that is, the ratio of assumed process error to measurement error, and also on the sample size n.

Table 1 : Chi-squared test: response of producer's risk to given increases to repeatability standard deviation, initial producer's risk 0.05

	Initial ratio of assumed process sd to repeatability sd			
Sample Size	1	$\sqrt{2}$	2	$2\sqrt{2}$
	Increase of 14% in repeatability sd			
5	0.083	0.071	0.062	0.057
10	0.099	0.081	0.068	0.059
15	0.112	0.089	0.072	0.062
20	0.124	0.096	0.075	0.063
25	0.135	0.102	0.079	0.065
30	0.146	0.108	0.081	0.066
	Increase of 10% in repeatability sd			
5	0.072	0.065	0.059	0.055
10	0.083	0.071	0.062	0.057
15	0.091	0.076	0.065	0.058
20	0.098	0.080	0.067	0.059
25	0.105	0.084	0.069	0.060
30	0.111	0.088	0.071	0.061

The table above gives the producers risk α resulting from increases in the repeatability standard deviation of 14% and 10%. The producer's risk before the increase is 0.05, and the various columns are for various ratios of process to repeatability standard deviations. The table as given applies to the case where duplicates are not used (d=1). The effect of using duplicates (d=2) is to move one column to the right. It should be noted that, although a ratio of process to measurement error of 3 is often taken as a point beyond which measurement error need not be considered, the measurement error in this context is taken as reproducibility.

Judging by the table, an increase of 10% in repeatability standard deviation would be unsatisfactory until the initial ratio of standard deviations reaches about 1.5, and an increase of 14% may be marginally acceptable when the initial ratio is 2, if we are prepared to accept an increase in producer's risk form 0.05 to 0.075.

The value of using duplicates in this test is clear from the table.

The relevance of the table is made clear by noting that the fat content of New Zealand butter from one factory is currently tested by EU customs for conformity with a process standard deviation of 0.14 using a method with a repeatability of 0.08. However, failure does not result in automatic rejection of the lot, and it is thought that this will usually be the case.

CONCLUSIONS SO FAR

Apart from method bias, which is considered later, it seems clear that the things to control are σ_L and σ_r, and that these need to be controlled separately. On the basis that an increase in producer's risk from 0.05 to 0.075 could be considered tolerable, we have suggested that increases in either or both these quantities of up to 14% could be acceptable, although with serious reservations concerning the effect on test 0 (several samples against a cut-off) and some reservations about the chi-squared test in 0. It seems likely that an increase of 28% would not be considered acceptable.

However, it is one thing to estimate that σ_L has not increased by 14% and quite another thing to prove it. Simple estimates suggest that an estimate of σ_L from a trial using the minimum of eight laboratories could be out by a factor of up to 50% either way. Even in cases where the Horwitz equation is considered appropriate, it may be questioned whether it is capable of pinning down σ_L to within 14%. The question of onus of proof must be addressed.

THE ONUS OF PROOF – METHOD VALIDATION AS ACCEPTANCE SAMPLING

The testing of a method for acceptability has on the conceptual level a strong resemblance to acceptance sampling. We may think of the proponents of the method as the "producer" and the regulatory authority to whom the method is proposed as the "consumer." The performance of the currently prescribed method sets a specification limit, and compliance (acceptability of the method) is determined by taking a random sample of the laboratories by which the method is proposed to be used. Key statistics from this sample are the mean (bias) and standard deviation (reproducibility, or something similar), which are subject to both sampling and measurement error.

From this point of view it is apparent that, as with all acceptance sampling, the problem is one of balancing the producer's and consumer's risks. In acceptance sampling this problem is normally dealt with, at least in theory, by taking a sample of sufficient size to render both risks acceptable, but in the method validation context the sample size (the number of laboratories involved in the validation trial) may be severely constrained.

Consequently we may expect that at least one of the parties will have to accept a significant amount of risk. If there is to be only a small risk of rejecting a candidate method that in truth performs as well or better than the currently prescribed method, we may expect a substantial probability that a relatively poor method will be accepted. If there is to be a high chance of rejecting inferior methods, then there is also a good chance of a satisfactory method being rejected.

RISKS FOR "ACCEPTANCE SAMPLING" OF METHODS

TRIAL SET UP AND THE "EFFECTIVE NUMBER OF LABORATORIES"

A trial is envisaged in which each of a number of laboratories test each of a number of samples, and that happens on a number of occasions (runs.) In such circumstances the repeatability standard deviation will be fairly well estimated, and the trouble comes with σ_L. This can be considered as having two components,

$$\sigma_L^2 = \sigma_{lab}^2 + \sigma_{run}^2,$$

and the separation is not well defined. Depending on how varied the conditions are within laboratories and between runs, variation will move from σ_{run}^2 to σ_{lab}^2 or in the reverse direction, σ_L^2 remaining constant. If the number of occasions is reasonable, σ_{run}^2 is relatively well estimated compared to σ_{lab}^2. For good estimation of σ_L^2 it is desirable, by varying conditions as much as possible between occasions within laboratories, to move as much variation as possible out of the poorly estimated σ_{lab}^2 into the well estimated σ_{run}^2. Looked at in another way, we are attempting to use different runs within a laboratory to some extent as if they came from different laboratories, thus increasing the effective

number of laboratories in the trial.

The extent to which this has been successful will be seen when the relevant analysis of variance is performed. Although the calculations involved are fairly complicated and approximate, procedures are available to approximate the resulting estimate of σ_L^2 by a multiple of a chi-squared variate, and the number of degrees of freedom attached to this estimate, plus one, could be considered as the effective number of laboratories involved in the trial. It would be expected to lie somewhere between the actual number of laboratories and the total (over all laboratories) number of occasions. Unless the number of available laboratories involved is large, it will be necessary for this procedure to be reasonably successful to give a reasonable chance of acceptance of the method according to the criteria suggested. In discussing what this chance might be, we shall use $nlabs'$ to denote the effective number of laboratories, and assume that the estimate s_L^2 has the distribution implied by this chi-squared approximation.

PROBABILITIES OF METHOD ACCEPTANCE

Consider the consequences of requiring a 90% upper confidence limit for σ_L to be less than 1.14 times σ_L for the standard method. This leads to the following table, which gives the chance that a method will be accepted.

Table 2: Probabilities of acceptance when a 90% upper confidence limit is required.

$nlabs'$	Ratio of σ_L (candidate method) to σ_L (standard method)						
	0.86	0.93	1	1.07	1.14	1.21	1.28
8	0.34	0.25	0.18	0.14	0.10	0.07	0.06
12	0.45	0.32	0.22	0.15	0.10	0.07	0.04
16	0.55	0.39	0.26	0.16	0.10	0.06	0.04
20	0.63	0.44	0.29	0.17	0.10	0.06	0.03
24	0.70	0.50	0.32	0.18	0.10	0.05	0.03
28	0.76	0.55	0.34	0.19	0.10	0.05	0.02
32	0.81	0.59	0.37	0.20	0.10	0.05	0.02
36	0.85	0.63	0.40	0.21	0.10	0.04	0.02
40	0.88	0.67	0.42	0.22	0.10	0.04	0.02
44	0.91	0.71	0.45	0.23	0.10	0.04	0.01
48	0.93	0.74	0.47	0.24	0.10	0.04	0.01
52	0.94	0.77	0.49	0.25	0.10	0.03	0.01
56	0.95	0.79	0.51	0.25	0.10	0.03	0.01
60	0.96	0.81	0.53	0.26	0.10	0.03	0.01

It will be seen that this criterion is quite good at rejecting poor methods, but also quite good at rejecting good ones. Until the effective number of laboratories rises to 50 a method has to actually improve on the standard method to have a 50/50 chance of acceptance.

If we relax the criterion to an 80% bound, the situation looks more somewhat more promising for candidate methods.

Table 3: Probabilities of acceptance when an upper 80% confidence limit is required

nlabs'	Ratio of σ_L (candidate method) to σ_L (standard method)						
	0.86	0.93	1	1.07	1.14	1.21	1.28
8	0.54	0.43	0.34	0.26	0.20	0.15	0.12
12	0.66	0.51	0.39	0.28	0.20	0.14	0.10
16	0.74	0.58	0.43	0.30	0.20	0.13	0.08
20	0.81	0.64	0.47	0.31	0.20	0.12	0.07
24	0.86	0.69	0.50	0.33	0.20	0.11	0.06
28	0.89	0.73	0.53	0.34	0.20	0.11	0.06
32	0.92	0.77	0.56	0.35	0.20	0.10	0.05
36	0.94	0.80	0.59	0.37	0.20	0.10	0.04
40	0.96	0.83	0.61	0.38	0.20	0.09	0.04
44	0.97	0.85	0.63	0.39	0.20	0.09	0.04
48	0.98	0.87	0.66	0.40	0.20	0.08	0.03
52	0.98	0.89	0.68	0.41	0.20	0.08	0.03
56	0.99	0.91	0.70	0.42	0.20	0.08	0.03
60	0.99	0.92	0.71	0.43	0.20	0.07	0.02

A method now has at least a 50/50 chance of passing once the effective number of laboratories gets above 24, which may well be achievable, provided that it is at least as good as the standard method. But 50/50 chance of acceptance would probably not be considered much of a reward for a great deal of work. The entries near the top right hand corner are not very satisfactory, given that we have decided that a 28% increase is not acceptable, and it may not be advisable to relax the criterion further.

TREATMENT OF THE METHOD BIAS

The uncertainty surrounding an estimate of method bias needs to be taken into account.

While for some purposes it may be appropriate to combine this uncertainty with σ_L to give an overall uncertainty, this conceals a fundamental difference between the two uncertainties: σ_L describes random variation, whereas the standard error of the bias σ_b does not. Conceptually a single random variable with standard deviation σ_b is generated when the validation trial is conducted, and thereafter applies without change whenever the method is used. Whereas the producer's risk will increase with high values of the random variable described by σ_L, the risk will decrease with low values, and at least over the long term may be expected to average out. With σ_b this averaging does not occur. Once you get an unlucky value, you are stuck with it.

It is therefore suggested that in cases where the uncertainty surrounding the bias is not negligible, a separate tolerance should be prescribed to allow for it. This could be in the form of a 95% confidence interval for the bias, with the upper limit being used as an additional tolerance when testing against an upper limit, and the lower limit being used when testing as a lower limit. When only one limit is applicable, an upper or lower one-sided interval could be used. The overall effect of this is roughly to use a tolerance of $k_a\sigma_b + k_a\sigma_L$ instead of the smaller value $k_a\sqrt{\sigma_b^2 + \sigma_L^2}$ that would be suggested if

the uncertainties were combined: it is as if σ_b were being added to σ_L.

If this point of view is accepted, the conditions under which σ_b can be considered negligible become considerably more stringent, and failure to allow for a σ_b equal to 14% of σ_L would have the same potential impact on producer's risk as the 14% "permissible" increase in σ_L already discussed, that is to increase it from 0.05 to 0.075. In fact, a σ_b as small as this could be realistically expected only from a trial involving at least 50 effective laboratories, with no contribution from uncertainty in reference values. The combined increase in risk would probably be considered unacceptable. There is a possibility of sharing the risk, say by dividing the 14% increase into two parts, one for a possible bias and the other for the method variance parameters, but it is thought that the increased stringency required for the estimation of the variance parameters would probably push the whole scheme outside the realms of practical feasibility, if it is not outside them already.

Accordingly, it is recommended that one of the criteria for acceptance of a method should be the provision of one-sided upper and lower 95% confidence limits for the method bias, and a requirement to allow appropriate tolerances, in addition to any others that may be required, when the method is used.

ANNEX C: EXAMPLES OF USE OF METHODS DESCRIBED IN ANNEX A

INTRODUCTION

The purpose of this Annex is to illustrate the methods of calculation involved in the analysis of a trial according to the methods suggested in Annex A, particularly those that involve more advanced statistical methods.

The first three illustrate various parts of the analysis of a single trial. This is described immediately below. The fourth example is set in a similar context, but introduces a variation in the trial design described there. In all, the examples seem to make it clear that a statistically skilled analyst will be required, not only to carry out the calculations correctly, but also to give confidence that the calculations are in fact those appropriate to the context and trial design.

Examples 1 to 3 are based on the assessment of the fat content of butter. This is required to lie between 80% and 82%. The data are not from a real trial but are obtained by simulation. It is assumed that a standard method with a repeatability standard deviation 0.080 pp and reproducibility standard deviation 0.160 pp is normally prescribed, and that the candidate method is being tested as a general replacement for this. These same values were also used to simulate results for the candidate method, which was also assumed to be unbiased: in fact the candidate and standard methods are equivalent.

The trial design was as follows.

> Five samples were presented to each of 10 laboratories, twice each as blind duplicates, over a period of 10 months. The samples were to be analysed under (within-laboratory) reproducibility conditions, with a separate run for each presentation. For each presentation duplicate analyses were be performed under repeatability conditions. This gives (10 laboratories).(10 presentations).(2 repeatability duplicates) = 200 analyses.
>
> Certified reference material was not used, but as part of the trial the samples were also analysed twice using the standard method, by each of four laboratories.

Use of this design in the examples does not constitute a recommendation for its general use. In particular, more precise estimation of the "reference" values under the standard method would be desirable. The range of reference values used is also barely adequate. This reflects a real life difficulty: it is in fact very difficult to obtain butter that is significantly outside product specification limits.

The units in which the data are expressed are percentages by weight.

EXAMPLE 1: ESTIMATION OF CONFIDENCE LIMITS FOR PRECISION PARAMETERS USING ANALYSIS OF VARIANCE.

Step 1

The design is suitable for analysis of variance. In the example, the analysis of variance came out as follows.

Table 4: Analysis of variance

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	Notation
Between Laboratories	2.33864	9	0.25985	A
Between Runs within Laboratories				
Between Samples	16.0873	4	4.02183	
Residual	2.60016	86	0.03023	B
Between repeatability duplicates	0.65190	100	0.00652	C
Total	21.67801	199		

Note 1

The entries in the table can be computed from the raw data as described below. However, it will generally be preferable to use a good statistical package to obtain the estimates, as allowance must often be made for missing data due to the discarding of outliers.

In the absence of such problems, the entries in the table may be obtained as follows:

Let *T* be the total, *M* the mean of all 200 results.

Let T_{lab} be the total, M_{lab} be the mean, of the 20 results for each laboratory *lab*. Then the between-laboratories sum of squares is calculated as

$$SS_{labs} = \sum T_{lab} M_{lab} - TM$$

with the sum being taken over all 10 laboratories. The number of degrees of freedom is one for each term in the sum, less one for the correction term TM, that is, nine.

A similar calculation using the total and mean of the 40 results for each sample yields the between-samples sum of squares $SS_{samples}$.

Now let SS_{runs}, with 99 degrees of freedom be obtained using a similar calculation using the total and mean of the 2 results for each laboratory in each run. From this is subtracted SS_{labs} and $SS_{samples}$ to give the residual sum of squares for runs within laboratories (line B in the table.) A similar calculation using degrees of freedom instead of sums of squares yields (99 – 9 – 4) = 86 degrees of freedom for this sum of squares.

Now the total sum of squares, with 199 degrees of freedom is calculated:

$$SS_{total} = \sum T_i M_i - TM$$

where T_i and M_i are now identical, each being taken over a single individual result i, there being 200 terms in the sum. From this is subtracted SS_{runs} with 99 degrees of freedom to give the sum of squares between repeatability duplicates.

Alternatively, the sum of squares between repeatability duplicates may be obtained as half the sum of the squared differences between repeatability duplicates.

Note 2

The 86 degrees of freedom available for estimating between run variation is more than the 50 degrees of freedom available from direct comparisons of reproducibility duplicates within laboratories. The contribution of these comparisons to the sum of squares can be computed directly, by summing the squared differences of the 50 pairs of results (averaged over repeatability duplicates) in which the same sample was analysed by the same laboratory. The additional 36 degrees of freedom come from comparisons in which the estimated differences between different samples are compared for different laboratories.

Step 2

An expression must now be obtained for the expectations of each of the mean squares A, B and C in terms of the variance components, σ^2_{lab} σ^2_{run} and σ^2_r. These expressions may often be obtained from the statistical package used to carry out the analysis of variance, and are as follows.

$$E(MS_A) = \sigma_r^2 + n_1\sigma_{run}^2 + n_1 n_2 \sigma_{lab}^2$$
$$E(MS_B) = \sigma_r^2 + n_1\sigma_{run}^2$$
$$E(MS_C) = \sigma_r^2$$

where n_1 is the number of repeatability duplicates per run (in this case 2) and n_2 is the number of runs per laboratory (in this case 10.)

Thus in the case being considered, we have

(1)
$$E(MS_A) = \sigma_r^2 + 2\sigma_{run}^2 + 20\sigma_{lab}^2$$
$$E(MS_B) = \sigma_r^2 + 2\sigma_{run}^2$$
$$E(MS_C) = \sigma_r^2$$

Step 3

We now express the precision parameters that we want to estimate in terms of the expected mean squares.

For example

(2a)
$$\begin{aligned}\sigma_L^2 &= \sigma_{lab}^2 + \sigma_{run}^2 \\ &= \frac{E(MS_A) - E(MS_B)}{20} + \frac{E(MS_B) - E(MS_C)}{2} \\ &= 0.05E(MS_A) + 0.45E(MS_B) - 0.5E(MS_C)\end{aligned}$$

and similarly

(2b)
$$\begin{aligned}\sigma_R^2 &= \sigma_{lab}^2 + \sigma_{run}^2 + \sigma_r^2 \\ &= 0.05E(MS_A) + 0.45E(MS_B) + 0.5E(MS_C)\end{aligned}$$

with of course

(2c)
$$\sigma_r^2 = E(MS_C).$$

To obtain estimates of these parameters we substitute for the expectations the observed values of the mean squares, to give

$\hat{\sigma}_L^2 = 0.05(0.25985) + 0.45(0.03023) - 0.5(0.00652) = 0.02334$, $\hat{\sigma}_L = 0.153$

and

$\hat{\sigma}_R^2 = 0.05(0.25985) + 0.45(0.03023) + 0.5(0.00652) = 0.02692$, $\hat{\sigma}_R = 0.164$

with

$\hat{\sigma}_r^2 = 0.00652$, $\hat{\sigma}_r = 0.081.$

Here we follow standard practice of using a caret to denote an estimate of the parameter concerned.

Step 4

The sampling variances (that is, the squares of the standard errors) of the independent mean squares A, B and C are now estimated, and estimates of the sampling variances of $\hat{\sigma}_L^2$ etc. are deduced from the equations in step 3.

Each mean square is proportional to a $\frac{\chi_\nu^2}{\nu}$ variate, where ν is the number of degrees of freedom, and so has coefficient of variation $\sqrt{\frac{2}{\nu}}$. This leads to the following estimates of the sampling variances.

$$\text{var}(MS_A) = \frac{2}{9}(0.25985)^2 = 1.500 \times 10^{-2}$$

$$\text{var}(MS_B) = \frac{2}{86}(0.03023)^2 = 2.125 \times 10^{-5}$$

$$\text{var}(MS_C) = \frac{2}{100}(0.00652)^2 = 8.50 \times 10^{-7}$$

We now use the method for combining variances of independent random variables,

$$\text{var}\left(\sum \alpha X\right) = \sum \alpha^2 \text{ var}(X)$$

to obtain the variances, and hence the standard errors, of the estimates obtained at step 3 from the equations (2a) and (2b) above.

$$\begin{aligned}\text{var}(\hat{\sigma}_L^2) &= 0.05^2 \text{ var}(MS_A) + 0.45^2 \text{ var}(MS_B) + (-0.5)^2 \text{ var}(MS_C) \\ &= 3.750 \times 10^{-5} + 4.303 \times 10^{-6} + 2.125 \times 10^{-7} \\ &= 4.202 \times 10^{-5}\end{aligned}$$

so that the standard error of $\hat{\sigma}_L^2$ is estimated as $\sqrt{4.202 \times 10^{-5}} = 0.00648$

$\text{var}(\hat{\sigma}_R^2)$ works out the same, since the formula is the same except for a change of sign in the coefficient of MS_C.

Step 5

We now obtain confidence intervals for the precision parameters. For this purpose we use $\frac{\chi_\nu^2}{\nu}$ approximations to the distributions of $\hat{\sigma}_L^2$ etc. We estimate their coefficients of variation *c* and deduce the number of degrees of freedom, $\nu = \frac{2}{c^2}$.

For σ_L^2 we have an estimate 0.02334 with standard error 0.00648 and thus $c = \frac{0.00648}{0.02334} = 0.278$, leading to $\nu = 25.9$. Corresponding figures for σ_R^2 are $c = 0.241$ and $\nu = 34.5$.

For σ_r^2 no calculation is needed: we already know that $\nu = 100$.

An upper 80% limit for the standard deviation σ, which may be any of σ_L, σ_R or σ_L is now given by $\hat{\sigma}\sqrt{\frac{\nu}{\chi^2_{0.80}}}$, where $\chi^2_{0.80}$ is the percentage point of the Chi-squared distribution with an upper tail probability of 80%. Using integer approximations to the degrees of freedom, we have:

Table 5: Estimates and upper 80% confidence bounds for precision parameters

	Estimate	DF	$\chi^2_{0.80}$	Upper 80 % limit
σ_L	0.153	26	19.82	0.175
σ_R	0.164	34	26.94	0.184
σ_r	0.0808	100	87.95	0.0862

The data used were from a simulation with $\sigma_r = 0.080$, $\sigma_{run} = 0.100$, $\sigma_{lab} = 0.096$. Thus the true values were $\sigma_L = 0.139$, $\sigma_R = 0.160$.

EXAMPLE 2: CALCULATION OF ADDITIONAL TOLERANCES FOR PRECISION PARAMETERS.

This continues example 1 above. Suppose that the standard method has accepted precision parameters

$\sigma_r = 0.080$, $\sigma_R = 0.160$. This gives

$$\sigma_L = \sqrt{0.160^2 - 0.080^2} = 0.139$$

for the standard method.

Application of the 14% criteria shows that for general acceptability, a candidate method should then have an 80% upper confidence bound less than $0.080 \times 1.14 = 0.091$ for σ_r and less than $0.139 \times 1.14 = 0.158$ for σ_L. Inspection of the table at the end of example 1 shows that the candidate method meets this requirement in respect of σ_r but not in respect of σ_L. An additional tolerance is then required when the method is used.

The tolerance will depend on the compliance test used. Suppose that the compliance test specifies that a composite sample shall be analysed in duplicate, and the average of the two duplicate analyses shall fall between 80.00 and 82.00 inclusive.

The average of the two duplicates is subject to a measurement error with standard deviation $\sigma_m = \sqrt{\sigma_L^2 + \frac{\sigma_r^2}{2}}$. For the standard method this works out to 0.1501. For the candidate method we may continue to take $\sigma_r = 0.080$, but we must take σ_L at its upper 80% confidence bound of 0.175, giving $\sigma_m = 0.1839$. A tolerance of $1.645 \times (0.1839 - 0.1501) = 0.056$ should be used at each end of the compliance range. (Here 1.645 is the upper 5% point of the normal distribution.)

Thus, leaving aside the question of method bias, product should be considered compliant if the mean of the two duplicate analyses falls within the range 79.944 to 82.056.

EXAMPLE 3: ESTIMATION OF BIAS AND ITS STANDARD ERROR, CALCULATION OF TOLERANCES TO ALLOW FOR POTENTIAL BIAS

We continue the example treated in examples 1 and 2. Suppose that official reference samples were not used, but that as part of the investigation the five samples were also analysed using the standard method by four laboratories (possibly different from the ones testing the candidate method.) The five samples were analysed under repeatability conditions twice, on separate runs, in each of the four laboratories. The average analyte levels found are given in the table below.

Table 6: Sample means for standard method

Sample	1	2	3	4	5
Mean	79.916	80.333	80.991	81.549	82.054

These have a mean of 80.969 and a standard deviation (using an *n* rather than an *n*-1 divisor) of 0.778.

It is first necessary to consider what precision can be attached to the overall mean of 80.969. The measurement error attached to this mean is the total of:

- the average of four laboratory errors of standard deviation σ_{lab}
- the average of eight run errors of standard deviation σ_{run}
- the average of forty repeatability errors of standard deviation σ_r.

It will therefore have standard deviation $s_{ref} = \sqrt{\frac{\sigma_{lab}^2}{4} + \frac{\sigma_{run}^2}{8} + \frac{\sigma_r^2}{40}}$, where, as in Annex A, the notation

s_{ref} is used to denote the standard deviation of the uncertainty regarding the average level of the "reference" samples. For the standard method we do not know σ_{lab} and σ_{run}, and the experiment in which the reference values were assigned is not of adequate size to estimate them. To be conservative we have to use

$$s_{ref} \le \sqrt{\frac{\sigma_{lab}^2 + \sigma_{run}^2}{4} + \frac{\sigma_r^2}{40}} = \sqrt{\frac{\sigma_L^2}{4} + \frac{\sigma_r^2}{40}},$$

and for the standard method this works out, using the accepted values from example 2, at

$$s_{ref} \le \sqrt{\frac{0.137^2}{4} + \frac{0.080^2}{40}} = 0.074\,1$$

Next we consider whether a reasonably unbiased estimate of sensitivity is possible. Here, as noted in Annex A, it is not the uncertainty in the absolute values of the concentrations, but the uncertainty in their differences, that is relevant. The laboratory and run errors included in the averages in the Table 6 are the same for each sample, and consequently the differences between the averages for different samples are affected only by repeatability error. Each difference is affected by the difference between two averages each of eight repeatability errors, and it is therefore reasonable to take σ_e^2 in the sensitivity section of Annex A as $\frac{\sigma_r^2}{8} = 8.0 \times 10^{-4}$. The bias in the regression coefficient will be, from the discussion of sensitivity in Annex A, of the order of $\beta \frac{\sigma_e^2}{\sigma_x^2} = \frac{8.0 \times 10^{-4}}{0.778^2} \beta = 0.0013\beta$ which can certainly be treated as negligible.

The advantage of knowing the details of how the reference values were determined should be noted: had we known only that each reference value had an uncertainty with standard deviation 0.074 we should have had to take $\sigma_e = 0.074$, and anticipated a possible bias of about 0.009β, although this would still probably be quite acceptable.

The following table, Table 7, gives the individual estimates of bias and the regression coefficients for each laboratory, together with their means and standard deviations.

The estimates of bias are formed by averaging the differences between a laboratory's results and the sample averages in Table 6. In the absence of missing data, this is of course equivalent to subtracting the mean of Table 6 from the relevant laboratory mean. The estimates of sensitivity are obtained as regression coefficients, using the values in Table 6 as the independent variable.

Table 7: Estimates of bias and sensitivity

Laboratory	Estimated Bias	Estimated Sensitivity
1	0.1405	0.9295
2	-0.0205	1.0155
3	-0.0435	1.0165
4	-0.1015	0.9037
5	-0.1185	0.9750
6	0.1805	1.0022
7	0.0325	1.0401
8	0.0695	0.9227
9	0.2135	0.9651

10	0.0285	1.0676
Mean	0.0381	0.9838
Standard Deviation	0.1080	0.0511

From this the overall method bias is estimated as (using the notation of Annex A) $b(x_0) = 0.0381$, and the method sensitivity as $s = 0.9838$. The overall method bias is taken as applying at the average concentration from Table 6, $x_0 = 80.969$. Standard errors for these estimates are obtained by dividing the standard deviations by the square root of the number of laboratories, $s_{b(x_0)} = 0.0342$ and $s_s = 0.0311$. It will be seen that neither the estimate of overall bias nor the difference of s from unity is statistically significant, making it plausible that the candidate method is in fact unbiased over the concentration range 80 – 82 relative to the standard method. Estimates of bias and their standard errors at various points in the range can be constructed using the formulae given in Annex A,

$b(x) = b(x_0) + (s-1)(x - x_0)$ and $s_{b(x)} = \sqrt{s^2_{b(x_0)} + (x - x_0)^2 s_s^2}$.

However, as noted in Annex A, the estimate of $s_{x(x_0)}$ needs to be adjusted to allow for uncertainty regarding the overall level of the reference samples. The adjusted value, using the formula given there, is

$$\begin{aligned} s_{b(x_0)} &= \sqrt{0.0342^2 + s_{ref}^2} \\ &= \sqrt{0.0342^2 + 0.0741^2} \\ &= 0.0816 \end{aligned}$$

The final formulae then become

$$b(x) = 0.0342 + 0.0162(x - 80.969)$$

$$\begin{aligned} s_{b(x)} &= \sqrt{0.0816^2 + 0.0311^2 (x - 80.969)^2} \\ &= \sqrt{0.00666 + 0.000967(x - 80.969)^2} \end{aligned}$$

which are tabulated in columns 2 and 3 of the table below. The required 95% one-sided upper and lower confidence limits for the bias are obtained from

$$b(x) \pm 1.645 s_{b(x)}$$

and are given in columns 4 and 5.

Table 8: Upper and lower confidence limits for bias

Concentration	Estimated bias	Standard error	Lower one-sided 95% confidence limit for bias	Upper one-sided 95% confidence limit for bias
80.0	0.0538	0.0830	-0.083	0.190
80.5	0.0457	0.0819	-0.089	0.180
81.0	0.0376	0.0815	-0.096	0.172
81.5	0.0295	0.0820	-0.105	0.164
82.0	0.0214	0.0832	-0.115	0.158

Thus a tolerance of 0.16 is required when testing against an upper limit of 82, and a tolerance of 0.08 is required when testing against a lower limit of 80.

It will be observed that in this case the largest contribution to the uncertainty comes from uncertainty related to the "reference" values of the samples used.

EXAMPLE 4: ESTIMATION OF CONFIDENCE LIMITS FOR PRECISION COMPONENTS USING A COVARIANCE MATRIX

This example uses a different experimental design. Again 5 samples were each presented twice, with repeatability duplicates analysed at each presentation, but this time the trial was compressed into five runs at each laboratory, as follows:

Run	1	2	3	4	5
Samples presented	1,3	1,4	2,3	2,5	4,5

The design is unbalanced, as not all differences between samples are estimated with equal precision. For example, samples 1 and 3 can be compared directly within runs, but samples 1 and 2 cannot. Thus analysis of variance cannot be used, and the data must be analysed as a mixed model using REML or a similar method. This will require a statistical computer package capable of handling general mixed models. Such a design would not have been used without taking statistical advice and ensuring that appropriate analysis tools were available.

The REML procedure of the statistical package Genstat produced the following output.

*** Estimated Variance Components ***

Random term	Component	S.e.
Lab	0.005693	0.003856
lab.run	0.010595	0.002720
lab.run.sample.dup	0.005776	0.000676

Interpreted, this means:

Table 9: Estimates of Precision Parameters

Parameter	Estimate	Standard Error of Estimate
σ^2_{lab}	0.005693	0.003856
σ^2_{run}	0.010595	0.002720
σ^2_r	0.005776	0.000676

The following covariance matrix for the estimates was obtained from the procedure.

Table 10: Covariances of Estimates of Precision Parameters

	σ^2_{lab}	σ^2_{run}	σ^2_r
σ^2_{lab}	1.487E-05		
σ^2_{run}	-1.474E-06	7.398E-06	

σ_r^2	2E-09	-1.22E-07	4.57E-07

The matrix is symmetrical and only the terms on and below the main diagonal are reported.

From Table 9 we obtain

$$\hat{\sigma}_L^2 = \hat{\sigma}_{lab}^2 + \hat{\sigma}_{run}^2 = 0.005693 + 0.010595 = 0.016288,\ \hat{\sigma}_L = 0.128$$

and

$$\hat{\sigma}_R^2 = \hat{\sigma}_{lab}^2 + \hat{\sigma}_{run}^2 + \hat{\sigma}_r^2 = 0.005693 + 0.10595 + 0.005776 = 0.022064,\ \hat{\sigma}_R = 0.149$$

The sampling variances are then obtained from Table 10 using

$$\begin{aligned}\text{var}(\hat{\sigma}_L^2) &= \text{var}(\hat{\sigma}_{lab}^2) + 2\,\text{cov}(\hat{\sigma}_{lab}^2, \hat{\sigma}_{run}^2) + \text{var}(\hat{\sigma}_{run}^2) \\ &= 1.487\times10^{-5} + 2(-1.474\times10^{-6}) + 7.398\times10^{-6} \\ &= 1.932\times10^{-5}\end{aligned}$$

and

$$\begin{aligned}\text{var}(\hat{\sigma}_R^2) &= \text{var}(\hat{\sigma}_{lab}^2) + 2\,\text{cov}(\hat{\sigma}_{lab}^2, \hat{\sigma}_{run}^2) + \text{var}(\hat{\sigma}_{run}^2) + 2\,\text{cov}(\hat{\sigma}_{lab}^2, \hat{\sigma}_r^2) + 2\,\text{cov}(\hat{\sigma}_{run}^2, \hat{\sigma}_r^2) + \text{var}(\hat{\sigma}_r^2) \\ &= 1.487\times10^{-5} + 2(-1.474\times10^{-6}) + 7.398\times10^{-6} + 2(2\times10^{-9}) + 2(-1.22\times10^{-7}) + 4.57\times10^{-7} \\ &= 1.954\times10^{-5}\end{aligned}$$

The standard errors of $\hat{\sigma}_L^2$ and σ_R^2 are then the respective square roots of these, namely 0.00440 and 0.00442. The standard error of $\hat{\sigma}_r^2$ is of course given directly in Table 9 as 0.000676.

We then proceed exactly as in Step5, Example 1 to obtain

	Estimate	Degrees of freedom	Upper 80% limit
σ_L	0.128	27.4, say 27	0.167
σ_R	0.149	49.8, say 50	0.180
σ_r	0.076	146	0.084